For this week, the chapter guides you in aligning sequences, using MUSCLE (and various permutations within it) and ClustalW. Follow along with the examples given, and duplicate them so that you have the experiences of doing each. The chapter also discusses alignment with GUIDANCE2.

Then, as your assignment to be turned in, continuing with last week's assignment, experiment with the alignment software on the sequences you downloaded last week. Align with MUSCLE, Muscle-fast, Muscle-prog, ClustalW, and GUIDANCE2 (using the MAFFT algorithm), noting in a separate text file your observations of any differences between the alignment methods. Unless you have a large number of sequences with a large number of sites to compare, you will probably not notice differences in speed, but you may notice differences in the output alignments.

Make sure that you have aligned by codon, to ensure the realism of the alignment, and make sure to trim off excess sequence if one or two sequences are clearly much longer than the rest due to a duplication or similar event (record any trimming in the text file above). Also make sure to eliminate duplicate sequences as discussed in the text (record any elimination in the text file above). Verify the you have sufficient p-distance (amino acid identity or sequence identity, depending on whether you are using coding sequences or not), and record your p-distance (and the converse identity value) in the text file you used above. Remember that the required p-distance varies according to whether you are using coding sequences or not.

Finally, in the text file, record your observations of the alignment(s). Discuss whether you had to trim sequences or eliminate duplicate sequences, and whether you had to remove any sequences because the p-distance did not meet your requirements. At this point, how do you feel about your chosen sequences and whether they will fulfill the purpose that you intended, when you chose them last week.

Submit a .mas file of your alignment here. Also paste the contents of the text file from above into the Comments area where you submit the assignment, so I can read it without having to open a document.

My first alignment using MUSCLE and default parameters showed consistent gaps near the beginning and the end of all of my sequences. The last two sequences added: Camelus ferus and Myotis brandtii have large gaps through the entire middle section of the sequences. When aligning via MUSCLE and codons, the alignments look rather similar. Part of me wonders if there is much of a difference between the alignment types as I am using MUSCLEX and not MUSCLE7 from the textbook. Aligning using ClustalW took SIGNIFICANTLY longer than MUSCLE (about 5-8 seconds compared to ~30 seconds). The ClustalW alignment that I got was significantly different from my MUSCLE alignment in regards to the last two sequences that I pointed out earlier. The large gaps in these sequences were not in the middle of the alignment, but rather near the end of the alignments. However, when I aligned it using ClustalW and the penalty setting recommended in the textbook, the alignment looked similar to the MUSCLE alignment. I had issues using MEGAX and figuring out how to change the presets to MUSCLE fast and MUSCLE-prog. After my alignment, I decided to not trim any sequences as I had no duplicates and I also did not have any overly long sequences.

Determining the p-distance for nucleotides using my MUSCLE alignment, the settings I used were: Variance Estimation Method -> None, Substitutions Type -> Nucleotide, Model/Method -> p-distance, Substitutions to Include -> d: Transitions + Transversions, Rates among Sites -> Uniform

Rates, Gaps/Missing Data Treatment -> Pairwise deletion. The p-distance for all of my sequences averaged around 0.1 to 0.2, with a few at 0.3. There was one sequence with a p-distance of 0.03, and upon closer investigation it was between Mus musculus and Mus pahari, both being from the same species. With that being said, I chose to delete the Mus pahari sequence as it is not necessary to have two sequences from such similar species. The overall average p-distance of my MUSCLE alignment (using amino acid substitution type) was 0.2616, which translates to 73.84% identity in my sequences chosen for alignment, well above the required 30%.

Using GUIDANCE2, I aligned my sequences using the MAFFT algorithm and received a GUIDANCE alignment score of 0.980491. Visually inspecting the color-coded MSA, I found a sequence from about 53-150 that displayed low levels of GUIDANCE scoring as well as a section near the end of my sequences around 937-979. However, the vast majority of alignment in my sequences scored highly.

The only sequence I had to remove was the very similar sequence of Mus pahari (which was similar to Mus musculus) as it had an identity of 0.03 (rather close to 0). Other than that, it appears that all of my sequences showed sufficient identity in my finished alignment. I believe that all of my sequences will help me in viewing the evolution of the APOE gene through different species as well as determining if there is a possibility of a model organism with which APOE can be studied more effectively.